词向量与 LDA 相融合的短文本分类方法*

张 群 王红军 王伦文
(中国人民解放军电子工程学院 合肥 230037)

摘要: 【目的】针对短文本主题聚焦性差以及严重的特征稀疏问题, 设计一种基于词向量与 LDA 主题模型相融合的短文本分类方法。【方法】从"词"粒度及"文本"粒度层面同时对短文本进行精细语义建模, 首先基于 Word2Vec 训练词向量并通过相加平均法合成"词"粒度层面的短文本向量, 基于吉布斯采样法训练 LDA 主题模型并根据主题概率最大原则对短文本进行特征扩展, 然后基于词向量相似度计算扩展特征权重得到"文本"粒度层面的短文本向量, 最后通过向量拼接构建词向量与 LDA 相融合的短文本表示模型, 在此基础上通过最近邻分类算法完成短文本分类。【结果】相比传统的基于向量空间模型、基于词向量、基于 LDA 主题模型这三种基于单一模型的分类方法, 词向量与 LDA 相融合的分类方法准确率、召回率、F_1 值均有提升, 分别至少提升 3.7%, 4.1%和 3.9%。【局限】仅应用于最近邻分类器, 尚未推广应用到朴素贝叶斯和支持向量机等多种不同的分类器。【结论】基于词向量与 LDA 相融合的短文本表示模型进行分类, 能有效克服短文本的主题聚焦性差及特征稀疏性问题, 提高短文本分类性能。

关键词: 短文本分类 词向量 LDA 主题模型 最近邻分类器
分类号: G350

1 引 言

移动终端的智能化催生了移动互联网的飞速发展。为适应移动用户阅读习惯, 移动互联网内容更多以短文本形式呈现, 例如微博和即时推送新闻等, 如何对海量短文本内容进行自动分类已成为研究者关注的热点问题。

在过去几十年里, 国内外学者提出及改进了一系列经典的机器学习算法, 如 k 近邻分类(k-Nearest Neighbors, k-NN) [1]、朴素贝叶斯分类(Naive Bayes, NB) [2]和支持向量机(Support Vector Machine, SVM)[3]等, 并将其成功应用于文本分类领域, 取得了比较满意的效果。然而相比普通长文本, 新兴的移动互联网短文本具有内容长度短小、信息描述能力弱、主题分散等特点, 使得以上经典文本分类方法应用于该领域时将面临严重的特征稀疏问题[4], 导致短文本分类效果并不理想。

文本数据表示对于文本分类至关重要, 数据表示的好坏直接影响分类效果。传统文本分类算法通常基于向量空间模型(Vector Space Model, VSM), 通过特征词及权值构成的向量表示文本数据[5]。该方法忽略了词语间的语义关系, 无法体现文本深层次的主题信息, 存在数据高维稀疏问题, 尤其是在表示短文本时, 语义缺失及高维稀疏问题变得更为严重。近年来针对这一问题的研究主要有三个方向。一些学者引入外部知识库(如搜索引擎、维基百科和知网等)对文本进行语义特征扩展以丰富词语间语义关系[6-7]。这些方法能一定程度上缓解稀疏性, 其局限性在于严重依赖外部知识库的质量, 对于知识库中未收录的主题概念无能为力, 且计算量大, 耗时长, 因此应用于主题分散的短文本效果一般。另有部分学者通过将原始高维特征词空间映射到低维的潜在语义空间或主题空间, 挖掘文本潜在的语义结构。如潜在语义分析方法(Latent Semantic Analysis, LSA)将文本表示为低维潜在语义

通讯作者: 张群, ORCID: 0000-0002-6196-7122, E-mail: zhangqunbit@163.com。
*本文系国家自然科学基金项目"动态数据挖掘的构造性机器学习方法研究"(项目编号: 61273302)的研究成果之一。

空间的语义向量[8]，降维去噪的同时改善稀疏性，但是降维过程可能带来分类受损问题且该语义空间每个维度的语义含义并不明确。相比LSA方法, LDA主题模型 (Latent Dirichlet Allocation, LDA)将文本表示为其隐含主题的概率分布[9]，能极大改善文本高维稀疏性，克服LSA方法分类受损问题的同时每个主题维度也具有可解释性，因此受到广泛应用。文献[10-11]直接在LDA主题维上进行文本分类，但由于短文本主题聚焦性差，该方法对于改善短文本的稀疏性效果有限；文献[12-14]基于 LDA 主题模型对短文本进行特征扩展，相比于单纯直接应用LDA的方法有一定的效果提升。以上的VSM、LSA和LDA模型均为直接导出短文本向量以表示短文本，属于"文本"粒度层面的模型。最新研究考虑从"词"粒度层面进行文本建模从而更精细地表达语义，首先导出词的向量表示，然后将词向量(Word Embedding)合成短文本向量[15]。这种方法有效解决了短文本主题分散和聚焦性差的问题，其局限性在于简单有效的词向量合成方法还有待研究，如文献[16-17]通过神经网络构建词向量的短文本合成模型，具有较高的复杂度。

在以上分析的基础上，本文将词向量与LDA有机融合，提出一种新的短文本分类方法，从"词"粒度及"文本"粒度层面同时进行短文本建模，以解决短文本特征稀疏问题及主题聚焦性差的问题。通过简单直接的相加平均法合成"词"粒度层面的短文本向量，避免了复杂的词向量合成过程；同时在进行"文本"粒度层面建模时，并非直接应用LDA模型将短文本映射到主题维，而是基于LDA主题概率最大原则对短文本进行特征扩展，并基于词向量相似度计算扩展特征权重，从而构建词向量与LDA相融合的短文本表示模型；另外在训练词向量及 LDA 模型时并不依赖已标注数据，仅在训练分类器时需要小规模的已标注数据，属于一种半监督学习方法[18]。

2 词向量训练及LDA建模

2.1 基于Word2Vec的词向量训练

词向量是词语的一种数学表示方法，向量的每个维度代表一个语义特征，向量间的距离或相似度能够反映词语间的语义相似性。分布假说理论(Distributional Hypothesis)表明词语语义由其上下文决定。依据分布假说理论，一种基于神经网络的词向量获取方法受到广泛研究，该方法通过对目标词的上下文及目标词与其上下文的关系进行建模，能够获取包含丰富语义的低维稠密的词向量。Bengio等提出神经网络语言模型(Neural Network Language Model, NNLM)，词向量作为一种副产品，是在训练该语言模型的同时得到的[19]。NNLM为一个三层前馈神经网络结构，如图1所示[19]。

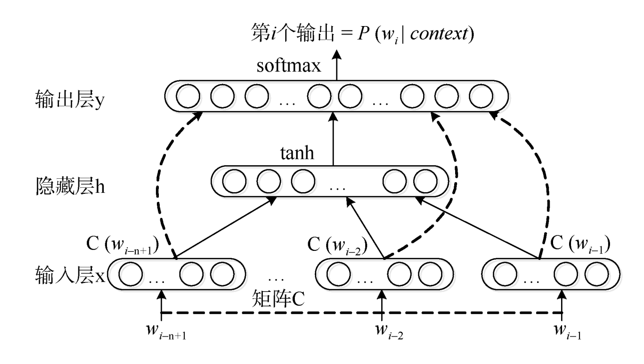


图 1 神经网络语言模型结构

NNLM结构图中，w_i为目标词，目标词的上下文为一个词序列，即 $context = \{w_{i-n+1}, \cdots, w_{i-2}, w_{i-1}\}$。NNLM的输入层通过一个矩阵**C**将上下文序列中的词映射为词向量，然后将词向量顺序拼接作为整个模型的输入，如下所示[19]。

$$\mathbf{x} = \{\mathbf{C}(w_{i-n+1}), \cdots, \mathbf{C}(w_{i-2}), \mathbf{C}(w_{i-1})\} \quad (1)$$

隐藏层与输出层分别如下[19]。

$$\mathbf{h} = \tanh(\mathbf{b} + \mathrm{H}\mathbf{x}) \quad (2)$$

$$\mathbf{y} = \mathbf{d} + \mathrm{W}\mathbf{x} + \mathrm{U}\mathbf{h} \quad (3)$$

其中，tanh为隐藏层激活函数，H为输入层到隐藏层的权重矩阵，U为隐藏层到输出层的权重矩阵，W为输入层到输出层的直连边权重矩阵(通常忽略)，**b**、**d**为模型偏置项。模型最终需通过Softmax函数将输出层**y**归一化为目标词的概率分布，如下所示[19]。

$$P(w_i \mid w_{i-n+1}, \cdots, w_{i-2}, w_{i-1}) = \frac{\exp(\mathbf{y}(w_i))}{\sum_{k=1}^{|\mathbf{V}|} \exp(\mathbf{y}(w_k))} \quad (4)$$

最后，模型通过迭代优化，在使公式(4)最大化的过程中训练出模型参数，其中包括词向量参数矩阵**C**，从而获得词向量。

NNLM 的计算量集中在公式(3)中的隐藏层到输出层的矩阵乘法U**h**中；另外，公式(4)中，|**V**|为词汇表大小，

因此当词汇表很大时 Softmax 函数计算非常耗时。

在 NNLM 的基础上，本文基于 Word2Vec 进行词向量训练。Word2Vec 是基于 Mikolov 等提出的 CBOW (Continuous Bag-of-Words)和 Skip-gram 模型开放的一款词向量训练工具[20]。CBOW 及 Skip-gram 这两种模型类似于 NNLM，区别在于 NNLM 是以训练语言模型为目标而间接获得了词向量，而 CBOW 和 Skip-gram 模型的直接目的即为获取词向量。因此 Word2Vec 在 NNLM 的基础上做了以下简化与改进:

(1) 去掉隐藏层，避免了公式(3)中复杂的矩阵乘法运算 **Uh**。

(2) NNLM 在输入层采用如公式(1)所示的词向量拼接法，而 Word2Vec 的 CBOW 模型采用向量相加求平均法降低了运算复杂度，如下所示[20]。

$$\mathbf{x} = \sum_{w_j \in c} \frac{\mathbf{C}(w_j)}{n-1} \tag{5}$$

其中，$c = \{w_{i-(n-1)/2}, \cdots, w_{i-1}, w_{i+1}, \cdots w_{i+(n-1)/2}\}$，指 CBOW 中目标词 w_i 前后各 $(n-1)/2$ 个词，即 w_i 的上下文。相比 NNLM 仅采用前 $(n-1)$ 个词作 w_i 的上下文，Word2Vec 更具有上下文完备性。

CBOW 与 Skip-gram 不同之处在于，CBOW 是通过上下文预测目标词而 Skip-gram 是通过目标词预测上下文。CBOW 与 Skip-gram 结构图分别如图 2、图 3 所示[20]。

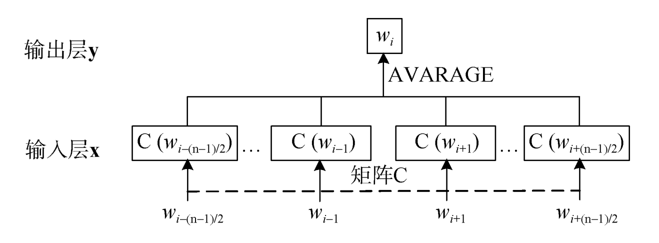


图 2 CBOW 模型结构

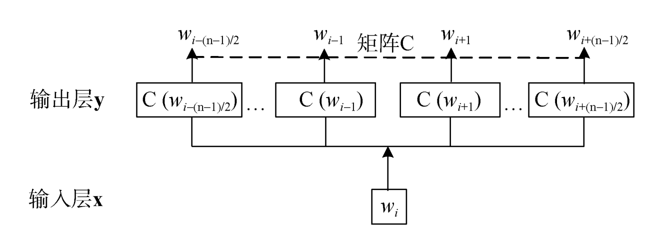


图 3 Skip-gram 模型结构

另外，针对 NNLM 输出层 Softmax 函数计算复杂度大的问题，Word2Vec 采用两种算法进行优化：结合霍夫曼编码的层次 Softmax 算法[21]及负采样(Negative Sampling)技术[15]。

本文基于 Word2Vec 训练词向量用于短文本分类任务，发现语料数据集规模及模型的选择会影响词向量质量进而影响短文本分类效果。针对这两个方面，总结以下经验用于指导训练词向量:

(1) 语料集规模在 200MB 以上时，CBOW 模型优于 Skip-gram 模型，在 100MB 以下则相反，在 100MB-200MB 之间两模型表现差别不明显。

(2) CBOW 模型在输入层采用词向量相加平均法代替 NNLM 中的词向量拼接法，降低了计算复杂度，但忽略了词序信息；本文尝试在 CBOW 模型的基础上仍采用词向量拼接法引入词序信息，但结果表明修改后的模型与原 CBOW 模型相比性能表现无明显差别。

2.2 基于吉布斯采样的 LDA 建模

LDA 主题模型是一个“文档–主题–词”的三层贝叶斯概率生成模型，其通过模拟文本的生成过程，将文本建模为混合主题上的概率分布，将主题建模为混合词上的概率分布，模型如图 4 所示[9]。

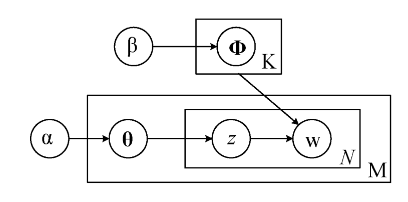


图 4 LDA 图模型

图 4 中符号含义如下：M 表示总文本数，N 表示一篇文本中的总词数，K 表示文本集隐含主题数；**θ** 为文本–主题分布矩阵，**Φ** 为主题–词分布矩阵，**θ** 与 **Φ** 均服从狄利克雷分布(Dirichlet Distribution)，α 为 **θ** 的超参数，β 为 **Φ** 的超参数；w 表示词，z 为 w 所属的主题。

令 $\mathbf{d}_m = (w_{m1}, w_{m2}, \cdots w_{mN})$ 表示第 m 篇文本，$\mathbf{z}_m = (z_{m1}, z_{m2}, \cdots, z_{mn})$ 中分量表示 $\mathbf{d}_m$ 中每个词对应所属的主题，$\mathbf{D} = (\mathbf{d}_1, \mathbf{d}_2, \cdots, \mathbf{d}_M)$ 表示整个文本集，$\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_M)$ 中分量表示 **D** 中每个文本对应的主题向量。基于图 4, LDA 模型生成过程描述如下:

(1) 对于第 m 篇文本 $\mathbf{d}_m$，根据 **θ** 服从参数为 α 的 Dirichlet 分布($\boldsymbol{\theta}_m \sim \mathrm{Dir}(\alpha)$)，确定一个主题分布 $\boldsymbol{\theta}_m$；

(2) 对于第 n 个词 w_{mn}，根据 z 服从 **θ** 的多项分布

($z_{mn} \sim Mult(\boldsymbol{\theta}_m)$)，为 w_{mn} 确定一个主题编号 z_{mn}；

(3) 根据 $\boldsymbol{\Phi}$ 服从参数为 β 的 Dirichlet 分布($\boldsymbol{\Phi}_m \sim Dir(\beta)$)，确定一个主题–词分布矩阵 $\boldsymbol{\Phi}_m$，同时根据步骤(2)确定的 z_{mn}，为 w_{mn} 确定一个词分布 $\boldsymbol{\Phi}_{z_{mn}}$；

(4) 根据词 w_{mn} 服从 $\boldsymbol{\Phi}_{z_{mn}}$ 的多项分布($w_{mn} \sim Mult(\boldsymbol{\Phi}_{z_{mn}})$)，生成词 w_{mn}；

(5) 遍历文本中 N 个词，重复步骤(2)–步骤(4)，生成 $\mathbf{d}_m$;

(6) 遍历文本集中M篇文本，重复步骤(1)–步骤(5)，生成整个文本集**D**。

LDA 模型的目标是为文本集 **D** 中的每个词分配一个潜在主题，从而估计出模型中的文本–主题分布矩阵 $\boldsymbol{\theta}$ 与主题–词分布矩阵 $\boldsymbol{\Phi}$，由此需要计算如公式(6)所示的后验概率[9]。

$$p(\mathbf{Z}|\mathbf{D}) = \frac{p(\mathbf{Z},\mathbf{D})}{\sum_{z} p(\mathbf{Z},\mathbf{D})} \tag{6}$$

其中分母计算难度非常大，为避免直接计算公式(6)，一种简单有效的方法是采用吉布斯采样(Gibbs Sampling)算法。

吉布斯采样算法[22]是一种特殊的基于马氏链的蒙特卡洛方法(Markov Chain Monte Carlo, MCMC)，通过对词的主题采样生成马氏链，用 $p(z_i|\mathbf{z}_{\neg i},\mathbf{D})$ 仿真近似 $p(\mathbf{Z}|\mathbf{D})$。$p(z_i|\mathbf{z}_{\neg i},\mathbf{D})$ 表示对于词汇表中 **V** 的一个词t，其当前采样的主题 z_i 依赖于其他时刻采样的主题 $\mathbf{z}_{\neg i}$。$p(z_i|\mathbf{z}_{\neg i},\mathbf{D})$ 通过吉布斯采样公式得到[22]。

$$p(z_i = k|\mathbf{z}_{\neg i},\mathbf{D}) = \frac{n_{k,\neg i}^{(t)} + \beta}{\left[\sum_{t=1}^{|V|}(n_k^{(t)} + \beta)\right] - 1} \cdot \frac{n_{m,\neg i}^{(k)} + \alpha}{\left[\sum_{k=1}^{K}(n_m^{(k)} + \alpha)\right] - 1} \tag{7}$$

其中，|**V**|表示词汇表 **V** 的大小；$n_k^{(t)}$ 表示词t采样为主题k的总次数，$n_{k,\neg i}^{(t)}$ 表示词t在其他时刻采样为主题k的次数；$n_m^{(k)}$ 表示文本 $\mathbf{d}_m$ 中采样为主题 k 的总词数，$n_{m,\neg i}^{(k)}$ 表示文本 $\mathbf{d}_m$ 中在其他时刻采样为主题k的词数。

主题采样完成后，基于采样得到的样本可以估计出模型的文本–主题分布矩阵 $\boldsymbol{\theta}$ 及主题–词分布矩阵 $\boldsymbol{\Phi}$，公式如下[22]。

$$\boldsymbol{\theta}_{m,k} = \frac{n_m^{(k)} + \alpha}{\sum_{k=1}^{K}(n_m^{(k)} + \alpha)} \tag{8}$$

$$\boldsymbol{\Phi}_{k,t} = \frac{n_k^{(t)} + \beta}{\sum_{t=1}^{|V|}(n_k^{(t)} + \beta)} \tag{9}$$

3 词向量与 LDA 结合的半监督分类方法

3.1 方法流程架构描述

LDA 主题模型从“文本”粒度层面对文本建模，在传统长文本分类任务中取得不错效果，但应用于短文本分类时效果很差；词向量属于“词”粒度层面的模型，在词语的语义相似度计算方面表现优越，但应用于文本级别的语义表示还有待研究。短文本介于“词”粒度与“文本”粒度之间，鉴于此，本文提出一种词向量与 LDA 相融合的短文本分类方法，从“词”粒度层面与“文本”粒度层面同时对短文本建模；另外，词向量及 LDA 模型的训练是在大规模无标注数据集上完成的，仅分类器的训练需要小规模的已标注训练数据，属于半监督学习方法。方法流程如图 5 所示。

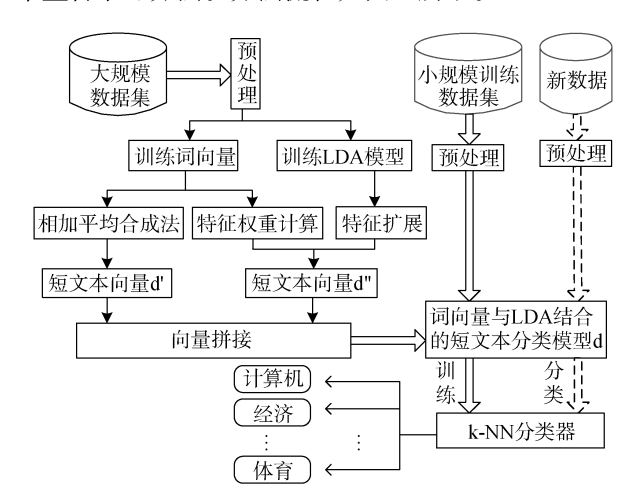


图 5　词向量与 LDA 融合的半监督分类方法流程架构

该方法分为 4 个步骤，描述如下：

(1) 构建一个大规模无标注数据集及一个小规模已标注数据集，并进行数据预处理；

(2) 在大规模无标注数据集上训练词向量及 LDA 主题模型；

(3) 在小规模已标注数据集上融合词向量与 LDA 对短文本建模；

(4) 构建一个 k 近邻分类器(k-NN)对新的短文本进行分类，测试本文方法的分类效果。

3.2 数据集构建及预处理

数据集的构建对于文本分类至关重要。分类任务属于有监督学习，需要大量已标注数据保证学习的准确性。本文分类方法属于半监督学习，仅需要小部分已标注数据，有效降低了人工数据标注的工作量。需

要构建一个大规模无标注数据集 **D** 及一个小规模有标注数据集 **D′**, 对两个数据集有以下要求:

(1) 数据集应符合正常的语言表达习惯;

(2) 数据集所包含的领域应与分类任务一致;

(3) 数据集应最大程度地包含并均衡分布于领域的各个潜在主题;

(4) 大规模无标注数据集应包含足够多的领域及主题相关词。

数据集预处理主要包括中文分词、停用词过滤等操作。对于小规模已标注训练数据集还需采用χ^2统计进行特征选择。χ^2统计值反映了词语 t 与数据集类别 c 的主题相关性, 如下所示[2]。

$$\chi^2(t,c)=\frac{(A\cdot D-B\cdot C)^2}{(A+B)\cdot(C+D)} \tag{10}$$

公式(10)中各参数含义如表 1 所示。

表 1 χ^2统计值参数含义表

	包含词语 t	不包含词语 t
属于 c 类	A	C
不属于 c 类	B	D

3.3 词向量与 LDA 融合的短文本表示模型

在大规模无标注数据集上训练词向量及 LDA 主题模型, 然后融合词向量与 LDA 对短文本建模。

基于 Word2Vec 训练词向量, 结果记为:

$$\mathbf{x}=\{\mathbf{C}(t_1),\mathbf{C}(t_2),\cdots,\mathbf{C}(t_{|V|})\} \tag{11}$$

其中, t_n表示词汇表 **V** 中第 n 个词, $\mathbf{C}(t_m)$为t_m的词向量表示。

基于吉布斯采样训练 LDA, 输出文件包括文本–主题分布矩阵$\boldsymbol{\theta}$、主题–词分布矩阵$\boldsymbol{\Phi}$及主题词文件。主题词文件显示了每个潜在主题下概率最大(即主题相关性最强)的前 n 个词, 主题词文件示例如表 2 所示。

表 2 主题词文件示例

主题编号	主题词及其概率值				
Topic 0th:	教育 0.020447	学校	0.017544	学生 0.015859	
	人 0.013244	教师	0.012354	……	
Topic 1th:	比赛 0.020663	中	0.012811	选手 0.011491	
	中国 0.011119	亚运会	0.010645	……	
Topic 2th:	中 0.009706	美国	0.007455	美军 0.006404	
	武器 0.006090	系统	0.006072	……	
Topic 3th:	软件 0.009364	函数	0.006048	系统 0.005572	
	程序 0.004344	过程	0.004271	……	

词向量与 LDA 结合的短文本建模方法具体实施步骤描述如下:

输入: ①小规模已标注的短文本训练数据集 **D′**;

②大规模无标注数据集 **D** 上训练得到的词向量;

③大规模无标注数据集 **D** 上训练得到的 LDA 模型。

输出: 训练数据集 **D′**的结合词向量与 LDA 的表示模型。

(1) 词向量合成

采用向量相加平均法得到 **D′** 的基于词向量合成的短文本表示模型, 如下所示。

$$\mathbf{d}_m{}'=\sum_{w_j\in\mathbf{d}_m{}'}\frac{\mathbf{C}(w_j)}{N_m} \tag{12}$$

其中, $\mathbf{d}_m{}'$表示 **D′** 中第 m 篇短文本的基于词向量合成的短文本表示, w_j为其中的词, N_m为词数, $\mathbf{C}(w_j)$为词w_j的词向量。

(2) 基于 LDA 进行特征扩展

将 **D′** 中的每个词与 LDA 模型的主题–词分布矩阵$\boldsymbol{\Phi}$相匹配, 选择该词所属的概率最大的主题z_{max}; 然后将z_{max}匹配LDA模型的主题词文件, 选择主题z_{max}下的前 r 个词作为该词的扩展特征, 则 **D′** 基于 LDA 的特征扩展模型如下:

$$\mathbf{d}_m{}''=\{w_{m1},(c_{11},c_{12},\cdots,c_{1r}),\cdots,w_{mn},(c_{n1},c_{n2},\cdots,c_{nr})\} \tag{13}$$

其中, $\mathbf{d}_m{}''$表示 **D′** 中第m篇短文本的基于LDA的特征扩展模型, w_{mn}为这篇短文本中的第 n 个词, $(c_{n1},c_{n2},\cdots,c_{nr})$为$w_{mn}$的 r 个扩展特征。

(3) 基于词向量的扩展特征权重计算

公式(13)中, 采用基于词频及逆向文档频(Term Frequency-Inverse Document Frequency, TFIDF)的方法计算被扩展特征w_{mn}的权重, TFIDF 权重反映了特征词表征文本的能力[1], 公式如下。

$$weight(w_{mn})=TFIDF(w_{mn})=\frac{TF(w_{mn})\cdot IDF(w_{mn})}{\sqrt{\sum_{w_{mn}\in\mathbf{d}_m''}[TF(w_{mn})\cdot IDF(w_{mn})]^2}} \tag{14}$$

其中, $TF(w_{mn})$表示w_{mn}的归一化词频, $IDF(w_{mn})$表示w_{mn}的逆向文档频, 分母部分是对 TFIDF 权重的归一化操作。

对于公式(13)中的扩展特征c_{nr}, 其权重与两个因素有关: c_{nr}所属的主题在文本中的重要性; c_{nr}与其所属主题的相关度。由于c_{nr}所属主题是由被扩展特征

w_{mn} 根据概率最大原则匹配 LDA 的主题–词分布矩阵得到的，因此认为 w_{mn} 的 TFIDF 权重代表 c_{nr} 所属的主题在文本中的重要性，c_{nr} 与 w_{mn} 的语义相关度代表 c_{nr} 与其所属主题的相关度。c_{nr} 与 w_{mn} 的语义相关度通过计算 c_{nr} 与 w_{mn} 的词向量的余弦值得到，记为 $sim(c_{nr},w_{mn})$，如下：

$$sim(c_{nr},w_{mn})=\frac{\mathbf{C}(c_{nr})\cdot\mathbf{C}(w_{mn})}{|\mathbf{C}(c_{nr})|\times|\mathbf{C}(w_{mn})|} \quad (15)$$

其中，$\mathbf{C}(c_{nr})$ 与 $\mathbf{C}(w_{mn})$ 分别为 c_{nr} 与 w_{mn} 的词向量表示。

综上，基于词向量的扩展特征权重计算方法如下：

$$weight(c_{nr})=TFIDF(w_{mn})\times sim(c_{nr},w_{mn}) \quad (16)$$

(4) 向量拼接

由于一个词可能含有多重语义，因此对于步骤(2)中的特征扩展模型 $\mathbf{d}_m''$，可能会出现同一扩展特征多次出现的情况，这时需合并相同的扩展特征，并将其权重相加作为合并后的扩展特征的权重。最终，将此特征扩展模型 $\mathbf{d}_m''$ 与步骤(1)中的基于词向量合成的模型 $\mathbf{d}_m'$ 进行顺序拼接，得到词向量与 LDA 结合的短文本表示模型，如下：

$$\mathbf{d}_m=\{\mathbf{d}_m';\mathbf{d}_m''\} \quad (17)$$

其中，“;”表示向量顺序拼接操作，$\mathbf{d}_m$ 为训练集 $\mathbf{D}'$ 中第 m 篇短文本的词向量与 LDA 结合的向量表示。

3.4 构建 k 近邻分类器

k 近邻分类(k-NN)算法作为一个经典的机器学习算法，应用于文本分类领域具有较高的稳定性，其原理简单直接：将新数据与训练数据集中的样本进行比较，选择与新数据最相似的前 k 个样本的类标签作为新数据的候选类标签，最后统计候选类标签中数量最多的类标签作为新数据的分类结果。

本文方法的最后一步通过构建一个 k-NN 分类器，以建模后的短文本训练集与待分类短文本数据作为输入，使用余弦相似度作为新数据与训练集样本的比较函数，完成对新数据的分类并测试本文方法的分类效果。

4 实验结果及分析

4.1 实验设置

采用复旦大学中文文本分类语料库作为大规模数据集用于训练词向量及 LDA 主题模型。选取 1 000 篇少于 150 字的短文本构建有类别标注的小规模训练数据集用于训练最近邻分类器，数据集均衡分布于计算机、经济、环境、艺术、体育 5 个领域，每个领域各 200 篇。另外选取 670 篇短文本作为测试数据集，其中，计算机类 145 篇，经济类 130 篇，环境类 135 篇，艺术类 120 篇，体育类 140 篇，训练集和测试集之间彼此不重叠，不包括重复文本。中文分词采用中国科学院计算技术研究所的 NLPIR 汉语分词系统。基于 Word2Vec 训练词向量，依据实验经验设置词向量维数为 50，当维数设置超过 50 时实验结果无明显提升。基于吉布斯采样方法训练 LDA 主题模型，依据 GibbsLDA++手册设置参数[23]，隐含主题数 K 设置为 100，超参数取 $\alpha=0.5$、$\beta=0.1$，主题词数设置为 20。依据实验经验设置 k-NN 分类器的近邻数 k，一般不超过训练样本数的平方根，取 k=20。

4.2 评价指标

分类结果用准确率(Precision, Pr)、召回率(Recall, Re)和调和平均值 F_1 三个指标来衡量，公式如下[1]。

$$Pr=\frac{TP}{TP+FP} \quad (18)$$

$$Re=\frac{TP}{TP+FN} \quad (19)$$

$$F1=\frac{2\cdot Pr\cdot Re}{Pr+Re} \quad (20)$$

各参数含义如表 3 所示。

表 3 分类评价指标参数含义表

	分类为 c 类	分类非 c 类
实际为 c 类	TP	FN
实际非 c 类	FP	TN

其中，准确率考察的是分类结果的正确性，召回率考察分类结果的完备性。

4.3 结果与分析

(1) 实验一

对 CBOW 模型进行修改，在 CBOW 模型输入层采用向量拼接法代替向量相加平均法引入词序信息，然后比较 Word2Vec 原版的 CBOW 与 Skip-gram 以及本文修改后的 CBOW 这三个模型训练的词向量应用于本文分类方法时所取得的分类效果，比较结果如表 4 所示。

表 4 词向量训练模型比较结果

词向量训练模型	准确率(%)	召回率(%)	F1 值(%)
Skip-gram	77.0	81.5	79.2
原版 CBOW(向量相加平均)	81.1	83.7	82.4
修改后 CBOW(向量拼接)	81.8	82.6	82.2

从表 4 可以看出：在短文本分类任务上，CBOW 模型优于 Skip-gram 模型；修改后的 CBOW 模型相比原 CBOW 模型仅在分类准确率上略有提升，而召回率及 F_1 值均略有下降，因此认为修改后的 CBOW 模型相比原 CBOW 模型无明显差别。综上，考虑到原 CBOW 模型计算复杂度低，因此本文方法基于原 CBOW 模型训练词向量。

(2) 实验二

测试本文方法在短文本分类任务上的分类效果，并与三种基于单一模型的分类方法(VSM+k-NN、词向量+k-NN、LDA+k-NN)进行比较，结果如表 5、表 6 所示。

表 5 本文方法分类效果

类别	准确率(%)	召回率(%)	F1 值(%)
计算机	85.3	87.1	86.2
经济	83.0	84.7	83.8
环境	79.3	84.2	81.7
艺术	78.3	80.6	79.4
体育	79.4	82.0	80.7
平均值	81.1	83.7	82.4

表 6 不同分类方法比较结果

分类方法	准确率(%)	召回率(%)	F1 值(%)
VSM+k-NN	74.7	77.2	75.9
词向量+k-NN	77.4	79.6	78.5
LDA+k-NN	66.2	69.3	67.7
本文方法 (词向量+LDA+k-NN)	81.1	83.7	82.4

表 5 显示本文分类方法在短文本数据集各个领域类别均能获得满意的分类效果，是一种有效的短文本分类方法。表 6 显示，前三种基于单一模型的分类方法中，基于 LDA 模型的方法分类效果最差，甚至低于传统的基于词袋模型的分类方法，表明 LDA 模型并不适用于短文本分类；与三种基于单一模型的分类方法相比，本文方法在三个分类指标上均有提升，其中分类准确率指标至少提升 3.7%，召回率至少提升 4.1%，F1 值至少提升 3.9%。这是因为方法融合词向量与 LDA 主题模型对短文本进行建模，能更精细地表示短文本语义信息，因此有效克服了单一 LDA 模型主题聚焦性差的缺陷以及词袋模型的特征稀疏问题，从而提高短文本分类效果。

5 结 语

本文提出一种同时从“词”粒度及“文本”粒度层面建模短文本的思路，并由此提出了一个词向量与 LDA 相融合的短文本分类模型。另外，该分类方法基于无标注数据集进行短文本建模，属于一种半监督学习方法。实验部分比较了该方法与三种传统基于单一模型方法的分类效果，此外还探讨了不同的词向量训练模型应用于本文方法时的优劣。后续将重点研究该分类方法应用于不同分类器的情况。

参考文献：

[1] Yang Y, Liu X. A Re-examination of Text Categorization Methods [C]. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2003:42-49.

[2] 邸鹏，段利国. 一种新型朴素贝叶斯文本分类算法[J]. 数据采集与处理，2014，29(1): 71-75. (Di Peng, Duan Liguo. New Naive Bayes Text Classification Algorithm [J]. Journal of Data Acquisition and Processing, 2014, 29(1): 71-75.)

[3] Joachims T. Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms [M]. Springer Berlin, 2002.

[4] 王仲远，程健鹏，王海勋，等. 短文本理解研究[J]. 计算机研究与发展，2016，53(2): 262-269. (Wang Zhongyuan, Cheng Jianpeng, Wang Haixun, et al. Short Text Understanding: A Survey [J]. Journal of Computer Research and Development, 2016, 53(2): 262-269.)

[5] Lebanon G. Metric Learning for Text Documents [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2006, 28(4): 497-508.

[6] 朱征宇，孙俊华. 改进的基于知网的词汇语义相似度计算[J]. 计算机应用，2013，33(8): 2276-2279,2288. (Zhu Zhengyu, Sun Junhua. Improved Vocabulary Semantic Similarity Calculation Based on HowNet [J]. Journal of Computer Applications, 2013, 33(8): 2276-2279,2288.)

[7] 王荣波，谌志群，周建政，等. 基于 Wikipedia 的短文本语义相关度计算方法[J]. 计算机应用与软件，2015，32(1):

82-85,92. (Wang Rongbo, Chen Zhiqun, Zhou Jianzheng, et al. Short Texts Semantic Relevance Computation Method Based on Wikipedia [J]. Computer Applications and Software, 2015, 32(1): 82-85, 92.)

[8] Deerwester S, Dumais S T, Furnas G W, et al. Indexing by Latent Semantic Analysis [J]. Journal of the Association for Information Science and Technology, 1990, 41(6): 391-407.

[9] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation [J]. Journal of Machine Learning Research, 2003, 3: 993-1022.

[10] 姚全珠, 宋志理, 彭程. 基于LDA模型的文本分类研究[J]. 计算机工程与应用, 2011, 47(13): 150-153. (Yao Quanzhu, Song Zhili, Peng Cheng. Research on Text Categorization Based on LDA [J]. Computer Engineering and Applications, 2011, 47(13): 150-153.)

[11] Rubin T N, Chambers A, Smyth P, et al. Statistical Topic Models for Multi-label Document Classification [J]. Machine Learning, 2012, 88(1-2): 157-208.

[12] 胡勇军, 江嘉欣, 常会友. 基于LDA高频词扩展的中文短文本分类[J]. 现代图书情报技术, 2013(6): 42-48. (Hu Yongjun, Jiang Jiaxin, Chang Huiyou. A New Method of Keywords Extraction for Chinese Short-text Classification [J]. New Technology of Library and Information Service, 2013(6): 42-48.)

[13] Chen M, Jin X, Shen D. Short Text Classification Improved by Learning Multi-granularity Topics [C]. In: Proceedings of the 22nd International Joint Conference on Artificial Intelligence. AAAI Press, 2011: 1776-1781.

[14] Phan X H, Nguyen L M, Horiguchi S. Learning to Classify Short and Sparse Text & Web with Hidden Topics from Large-scale Data Collections [C]. In: Proceedings of the 17th Information Conference on World Wide Web (WWW'08). New York: ACM, 2008:91-100.

[15] Mikolov T, Sutskever I, Chen K, et al. Distributed Representations of Words and Phrases and Their Compositionality [J]. Advances in Neural Information Processing Systems, 2013, 26: 3111-3119.

[16] Turney P D, Pantel P. From Frequency to Meaning: Vector Space Models of Semantics [J]. Journal of Artificial Intelligence Research, 2010, 37(1): 141-188.

[17] Kim Y. Convolutional Neural Networks for Sentence Classification [C]. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2014: 1746-1751.

[18] Chapelle O, Schlkopf B, Zien A. Semi-Supervised Learning [J]. Journal of the Royal Statistical Society, 2010, 6493(10): 2465-2472.

[19] Bengio Y, Ducharme R, Vincent P, et al. A Neural Probabilistic Language Model [J]. Journal of Machine Learning Research, 2003, 3(6): 1137-1155.

[20] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space[C]. In: Proceedings of Workshop at ICLR. 2013.

[21] Morin F, Bengio Y. Hierarchical Probabilistic Neural Network Language Model [C]. In: Proceedings of Workshop at AISTATS. 2005.

[22] Porteous I, Newman D, Ihler A, et al. Fast Collapsed Gibbs Sampling for Latent Dirichlet Allocation [C]. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Las Vegas, USA. 2008.

[23] GibbsLDA++: A C/C++ Implementation of Latent Dirichlet Allocation (LDA) Using Gibbs Sampling for Parameter Estimation and Inference [EB/OL]. [2016-05-15]. https://sourceforge.net/projects/jgibblda/.

作者贡献声明:

张群, 王红军: 提出研究思路, 设计研究方案;
张群: 进行实验, 采集、清洗和分析数据, 论文起草;
王红军, 王伦文: 论文最终版本修订。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据[1]由作者自存储, E-mail: 1875586718@qq.com; 支撑数据[2-4]见期刊网络版 http://www.infotech.ac.cn。

[1] 张群. dataset.zip. 用于训练词向量与 LDA 主题模型的大规模数据集.
[2] 张群. word2vec.zip. 用于训练词向量的 word2vec 源码.
[3] 张群. GibbsLDA++.zip. 基于吉布斯采样的 LDA 主题模型程序包.
[4] 张群. LDA_result.zip. LDA 主题模型训练结果文件.

收稿日期: 2016-08-01
收修改稿日期: 2016-10-14

Classifying Short Texts with Word Embedding and LDA Model

Zhang Qun Wang Hongjun Wang Lunwen
(Electronic Engineering Institute of PLA, Hefei 230037, China)

Abstract: [**Objective**] This paper proposes a short text classification method with the help of word embedding and LDA model, aiming to address the topic-focus and feature sparsity issues. [**Methods**] First, we built short text semantic models at the "word" and "text" levels. Second, we trained the word embedding with Word2Vec and created a short text vector at the "word" level. Third, we trained the LDA model with Gibbs sampling, and then expanded the feature of short texts in accordance with the maximum LDA topic probability. Fourth, we calculated the weight of expanded features based on word embedding similarity to obtain short text vector at the "text" level. Finally, we merged the "word" and "text" vectors to establish an integral short text vector and then generated their classification scheme with the k-Nearest Neighbors classifier. [**Results**] Compared to the traditional singleton-based methods, the precision, recall, F1 of the new method were increased by 3.7%, 4.1% and 3.9%, respectively. [**Limitations**] Our method was only examined with the k-Nearest Neighbors classifier. More research is needed to study its performance with other classifiers. [**Conclusions**] The proposed method could effectively improve the performance of short text classification systems.

Keywords: Short text classification Word embedding Latent Dirichlet Allocation k-Nearest Neighbors

哈佛大学图书馆选择 Ex Libris Alma 为其下一代图书馆平台

近日，哈佛大学图书馆已选择 Ex Libris Alma 图书馆管理服务作为该图书馆支持研究、教学和学习战略的一部分。哈佛大学是第 33 个选择 Alma 解决方案的研究图书馆协会(Association of Research Libraries, ARL)成员。

凭借其统一的资源管理功能、高级工作流程、强大的基础架构和云平台, Alma 解决方案将帮助哈佛大学图书馆实现在整个图书馆网络的单一框架内有效管理印刷和在线馆藏的目标。

哈佛大学图书馆馆长 Sarah Thomas 指出: "哈佛大学图书馆的战略目标之一是通过直观的发现系统、专业网络和全球合作，有效地访问知识和数据世界。Alma 是一个强大的平台，其强大的功能、易用性和云服务将帮助我们实现直接目标以及长期目标。"

Ex Libris 总裁 Matti Shem Tov 评论说: "哈佛大学图书馆是 Ex Libris 的长期合作伙伴，其自 2000 年以来一直在使用 Aleph ILS，自 2014 年以来一直使用 Primo 的发现和交付解决方案。哈佛大学图书馆采用 Alma 证明了 Alma 为全球顶级学术机构的图书馆提供管理服务的能力。我非常高兴哈佛大学图书馆现在加入了 Alma 社区，并期待其对这个活跃团体的贡献。"

(编译自: http://www.proquest.com/about/news/2016/Harvard-Library-Selects-the-Ex-Libris-Alma-Next-Generation-Library.html)

(本刊讯)